Automatic Tag Recommendation for the UN Humanitarian Data Exchange

Ghadeer Abuoda[a], Chad Hendrix[b] and Stuart Campo[b]

[a]College of Science and Engineering, Hamad Bin Khalifa University, Qatar
[b]United Nations Office for the Coordination of Humanitarian Affairs (OCHA), Centre for Humanitarian Data, Netherlands

Abstract

We have recently seen a rapid growth of data portals and dataset repositories being made available on the Web. While these repositories have been critical for advancing research, much work remains to improve finding appropriate datasets and relevant sources. Search engines, the primary tools for dataset discovery, are mainly keyword-based over published metadata of the datasets, whether within dataset repositories or over the Web. However, in most cases, the available metadata may not encompass the essential information the user needs to decide whether the dataset fits a given task. Therefore, data publishers should annotate their datasets with informative metadata when they add them to a dataset repository. Tags are a particular form of metadata that the data publisher uses to describe their view of how the dataset should be categorized. An interesting problem is how to automate the process of recommending tags to data publishers when they add new data to a dataset repository. In this paper, we develop an approach for automatic tag recommendation for dataset repositories. We investigate how to exploit the features of the dataset and the tagging history in the repository to build an effective tag recommendation model. We further demonstrate the integration of the model in the *The Humanitarian Data Exchange*, a real-world dataset repository in the social and humanitarian domains.

Keywords

Dataset Repository, Dataset Tagging, Keyword Search, Tag Recommendation

1. Introduction

Nowadays, many dataset repositories and data portals are created by different organizations to facilitate sharing and distribution of datasets. Online platforms like CKAN,[1] Quandl Kaggle,[2] and Microsoft Azure Marketplace[3] are examples of dataset repositories that host datasets for data-driven research in a wide range of domains. The data in these repositories is usually tabular (e.g., CSV files), and the goal of the repositories is to enable *data scientists* to find, access, integrate, and analyze combinations of datasets based on their needs. The first step in this process is to *find* the datasets relevant to a task, which requires *information retrieval*. Currently, dataset repositories use search engines that were mainly developed for unstructured textual documents. To improve retrieval quality, dataset repositories typically allow data publishers

BIRDS 2021: Bridging the Gap between Information Science, Information Retrieval and Data Science, March 19, 2021, online

gabuoda@hbku.edu.qa (G. Abuoda); hendrix@un.org (C. Hendrix); campo2@un.org (S. Campo)

[1]https://ckan.org/
[2]https://www.quandl.com/
[3]https://azuremarketplace.microsoft.com/en-us/marketplace/

to add *metadata* with their datasets, i.e., structured information about the data [1]. The search engines rely on this metadata in addition to the content of the datasets to guide the users toward relevant datasets. Thus, high-quality metadata plays an important role in enabling users to find datasets relevant to their needs [2, 3].

One type of metadata that data publishers often use to annotate and label their datasets is *tags* [4]. In particular, publishers can assign freely chosen keywords to datasets with the purpose of referencing these datasets later on with the help of these assigned tags. A dataset publisher can define their tags to describe a dataset as a whole or emphasize a certain topic that is only relevant to the dataset. A fundamental issue that underlies the effectiveness of user-defined tags is the quality and the relevance of these tags [5]. On the one hand, these tags represent a more flexible way of describing content than a fixed taxonomy with a controlled vocabulary, which means that tags should be freely chosen by data publishers. On the other hand, tags should be correctly formed and spelled, relevant to the content and its terms, and not repetitive or ambiguous. To balance these conflicting goals, a *tag recommendation* method can assist data publishers in the tagging process to improve the quality of the available metadata about their datasets [5]. Good tag recommendation can benefit not only search, but also other services that rely on tags such as content recommendation and categorization.

In this paper, we present a tag recommendation model and show how we applied it effectively to improve *information retrieval* for datasets in the *Humanitarian Data Exchange (HDX)* platform, in service of the *data scientists* who use this platform. The main idea of our model is to analyze the metadata and tagging history associated with existing datasets to find candidate tags. We propose a way of integrating the model in the dataset upload page, which encourages data publishers to attach informative tags to their dataset when they first upload them. Automatic tag recommendation raises user confidence when interacting with the platform: (i) dataset publishers feel more confident that they are not guessing how they can tag; the HDX platform makes them feel supported, and (ii) users who come to the HDX platform looking for information have more confidence because they have a more accurate picture of the datasets.

As an example of our model recommendations on the HDX platform, consider the dataset *Nigeria: 2018 Education Secondary Data Review (SDR)*[4] published by the Nigeria Education in Emergencies Working Group. The dataset contains assessment reports for humanitarian missions in the education domain. The dataset is currently tagged with only two tags, "EDUCATIONNEEDS" and "ASSESSMENT". Our model recommends additional tags and ranks them based on similarity to the dataset, with the top three tags being "Nigeria complex emergency", "education", and "education cluster". This gives the dataset publisher meaningful tagging options and higher confidence when tagging their dataset.

2. Related Work

In dataset repositories, tags form the source for enriching taxonomies in evolving and dynamic content published in these repositories [6]. Moreover, many metadata standards were developed to aid researchers in sharing research (data, code, publications) that rely on tagging techniques [7]. Additionally, dataset-centered search engines rely extensively on metadata

[4] https://data.humdata.org/dataset/nigeria-2018-education-secondary-data-review-sdr

generally and tags specifically for dataset discovery and retrieval. For instance, Google dataset search engine [3] crawls the web for all datasets and collect the associated metadata. These standards and tools are effective only if the metadata and tags are mainly correct and maintained. However, in practice, most datasets have incomplete or non-existent metadata [8]. Therefore, there is a need for work like ours to automate the creation of metadata.

Tag recommendation services have a direct benefit to IR services such as search [9] and query expansion [10]. There are many well-studied approaches for tag recommendation, such as content-based methods, collaborative filtering methods, and hybrid methods [5]. Regardless of the type of tag recommendation method, the challenge in tag recommendation is always in finding the appropriate set of tags that better describe a given resource.

Text analysis has long been recognized as a useful technique for extracting informative tags for web resources. In this approach, each resource (in our case a dataset) is represented as a document through a vector of all word occurrences weighted by term frequency-inverse document frequency (TF-IDF) or statistical topic modeling techniques. Various tag recommendation techniques have been proposed relying on different representations of the resources and computing the similarity between different resources in addition to mining the historical occurrence of tags [11, 12, 13, 14]. Next, we present how we use text analysis techniques to recommend tags in HDX.

3. Tagging on the Humanitarian Data Exchange (HDX)

The Humanitarian Data Exchange (HDX)[5] is an open platform for sharing data across crises and organizations. The HDX platform is managed by the Centre for Humanitarian Data of the United Nations Office for the Coordination of Humanitarian Affairs (OCHA). The platform hosts more than 17,000 datasets shared by hundreds of organizations covering humanitarian crises around the world. The goal of the HDX platform is to make humanitarian data easy to find and use for analysis. HDX platform has a search-engine interface that allows users to search datasets via keywords or a faceted search on features such as location, organization, licenses, etc. The returned datasets are presented in a structure-aware fashion, exposing attributes of the datasets (number of downloads, tags, dataset owner, format, etc.) and enabling users to explore different quick charts of the datasets or develop their own visualizations. A keyword search relying on user-generated metadata is the most common way to find a specific dataset on the platform. One crucial factor in defining the quality of the search results on the HDX platform is the quality and richness of the metadata, mainly the tags provided by dataset publishers.

On the HDX platform, the tag usually refers to a concept (e.g., health, education, camps), a specific crisis (e.g., Syria, Darfur), the type of the crisis (e.g., earthquake, hurricane), and/or the organization that collected the dataset (e.g., UNICEF, Education Above All). At the time of this work, there was no defined list of tags, and data publishers could use free text to tag datasets.

The HDX technical team reported that, in a particular sample of 19,171 search queries, only 8,114 resulted in actual downloads of HDX datasets. One possible interpretation of this gap between search requests and dataset downloads is that users may not be satisfied with the search results or could not find the information they expected. Since tags play an important

[5] https://data.humdata.org/

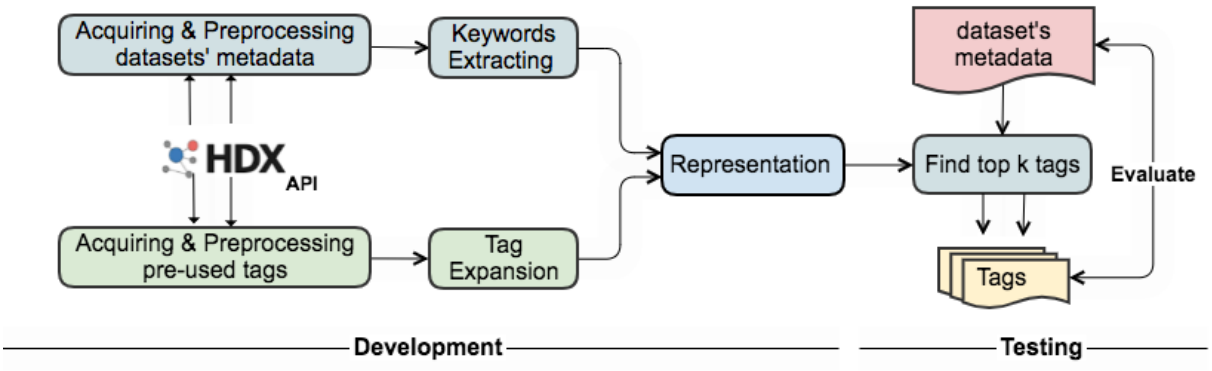

Figure 1: Phases of the HDX Tag Recommender

role in search quality, we propose a method for improving the tagging process on the HDX platform with the goal of improving search quality and user engagement.

4. Our Proposed Tag Recommender

Our model takes as input the set of tagged datasets in the repository, and an input target dataset d. The model should provide a list of top k candidate tags, sorted according to their relevance to dataset d. In this work, we investigate recommending tags that are relevant to target dataset d by utilizing various types of information: (i) previously assigned tags in the repository, (ii) terms extracted from textual features of the datasets in the repository (e.g., title, description, etc.), and (iii) terms extracted from the target dataset.

Developing our model on the HDX platform required us to address several challenges. First, the amount of metadata available varies widely between datasets. In some datasets, the metadata can contain around 1,000 different terms, while other datasets can barely reach 40 terms. Thus, our approach needs to enrich the metadata of datasets with a few terms by choosing the important words in the datasets' content. Second, data publishers use numbers, special characters, and hyperlinks in the description of their dataset. This content affects the ability to match with predefined tags and to define similarity in any approach (e.g., "Syria crisis-2011" is different from "Syria crisis"). Third, data publishers sometimes provide the description of their datasets in PDF files, not free text. In some cases, the attached PDF file reflects the project in which this dataset was collected, not a description of the dataset itself. Fourth, the tags used for HDX may refer to the same concept with different terms (e.g., education vs. learning; sex/age rate vs. demographics; displaced people location vs. displacement and shelter). Moreover, the valid list of tags contains more specific concepts (e.g., education in emergencies, education facilities). Fifth, data publishers use acronyms as tags. They use different acronyms to refer to the same concept (e.g., using both '3W' or '3Ws' to refer to a 'who-is-doing-what-where' dataset). Alternatively, they may use acronyms in a way that will change the meaning and make finding a match in the valid tag list even harder (e.g., using 'pin' to mean 'people in need'). Finally, the tags can be variations on the same term (e.g., refugee vs. refugees).

To address these challenges, we developed a model that analyzes the metadata of a dataset through different phases using off-the-shelf text processing techniques. The main phases of our recommendation model are depicted in Fig. 1, and are summarized in the rest of this section.

Acquiring Metadata Using the HDX API,[6] we extract the metadata collection associated with a group of datasets of interest. For example, we may be interested in the education domain and thus focus on datasets annotated with the "education" tag. The metadata elements extracted for each dataset are: the *title* of the dataset, the *tags* assigned to the dataset by the data publisher, the *organization* that provided this dataset, the *source* of the dataset is different from the organization, the *URL* which enables us to crawl the content of public datasets to enrich the metadata with information from the dataset header, the *countries* mentioned in the metadata object, whether the dataset has *geodata*, and the free-form *note* describing the dataset. The output of this phase is a record of terms extracted from the HDX metadata for each dataset.

Preprocessing and Cleaning This phase includes tokenization, stemming (e.g., "refugees" and "refugee" become the same token), and removing numbers/special characters/links/stopwords/non-English terms.

Candidate Tag Extraction The set of terms extracted from the metadata of the dataset collection is our *vocabulary*. We extract candidate tags from this vocabulary that could be an individual term or a pair of co-occurring terms. An important step in our work was to evaluate different methods for defining candidate tags (results in the next section). We evaluated: (1) scoring each vocabulary term based on Term Frequency (TF) and using the terms with the top TF scores as candidate tags, (2) combining TF with Inverse Document Frequency and using the top TF-IDF terms, and (3) extracting frequent co-occurring terms from the vocabulary using N-grams to help decide which N-terms can be chunked together to form a single tag.

Tag Expansion Our metadata acquisition step extracts the set of previously used tags in the repository. In the tag expansion phase, we enrich these tags by adding related terms. We expand by adding related terms from WordNet[7] (e.g., "teaching", "pedagogy", and "didactics" are added to the "education" tag). We also consider enriching the tags using similar terms based on the word2vec model [15].

Computing Similarity The model identifies a set of candidate tags from the vocabulary. It also uses a similar process to identify a set of candidate tags for the target dataset d. We need a similarity measure between the tags in the two sets. We explore different representation techniques such as vector encoding, TF, TF–IDF, and distributed representations (i.e., word2vec). We compute the *cosine similarity* between the representation of the candidate tags from the vocabulary and the target dataset.

Tags Recommendation We now reach the key step in our approach: recommending tags for the target dataset. Our model ranks the candidate tags by their similarity to the target dataset d and recommends the top k candidate tags.

Setting Thresholds We need to compute TF and TF–IDF thresholds to guide the creation of the vocabulary. These thresholds define the cut-off points that determine which terms to eliminate from the vocabulary. In order to determine the thresholds in our model, we test different cut-off

[6] https://github.com/OCHA-DAP/hdx-python-api
[7] https://wordnet.princeton.edu/

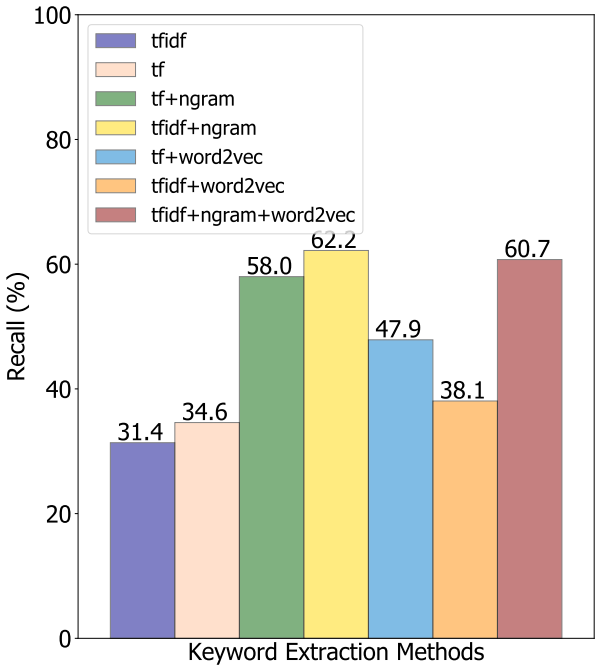

Figure 2: Recall for Different Keyword Extraction Methods

values and observe their effect on vocabulary size. There is typically a cut-off value where going higher leads to a significant reduction in vocabulary size, and this is the cut-off value we use.

5. Experimental Evaluation

Datasets and Tag Selection We used a subset of HDX datasets to develop and evaluate our model. There were around 3000 private and public datasets that are annotated with the tag "education". We sampled 80% of these datasets to build the vocabulary of the model while the remaining 20% were used for evaluating the recommended tags.

Evaluation Our model recommends k tags for each dataset (we set k in the range 3-5). Our evaluation metric is the percentage of these tags that is already used in tagging the dataset. This is a recall metric [16]. The vocabulary consists of around 1800 terms. Term frequency and document frequency vary widely, and thresholds TF=20 and DF=30% worked best. Fig. 2 shows the recall of different methods of building the vocabulary. Using frequency to identify candidate tags achieves around 30% recall. Adding N-grams (N=2) boosts recall by around 20 percentage points. Using word2vec is not effective, even when combined with N-grams. Thus, we recommend using TF-IDF and N-grams, but not the more complex word2vec.

6. Conclusion

We presented an approach to automatically recommend tags for datasets on the HDX platform. The effectiveness of our model lies in using existing metadata in the dataset repository in addition to the textual features of a dataset to recommend informative tags. Our goal is for better tags to lead to better search results and user engagement on the HDX platform.

References

[1] S. Khalsa, P. Cotroneo, M. Wu, A survey of current practice of data search services, 2018.

[2] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L.-D. Ibáñez, E. Kacprzak, P. Groth, Dataset search: a survey, The VLDB Journal 29 (2020).

[3] N. Noy, M. Burgess, D. Brickley, Google dataset search: Building a search engine for datasets in an open web ecosystem, 2019.

[4] P. Rafferty, Tagging, KO KNOWLEDGE ORGANIZATION 45 (2018).

[5] F. M. Belém, J. M. Almeida, M. A. Gonçalves, A survey on tag recommendation methods, Journal of the Association for Information Science and Technology (2017).

[6] F. Nargesian, K. Q. Pu, E. Zhu, B. Ghadiri Bashardoost, R. J. Miller, Organizing data lakes for navigation, in: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, 2020.

[7] P. Rocca-Serra, A. Gonzalez-Beltran, L. Ohno-Machado, G. Alter, The data tags suite (dats) model for discovering data access and use requirements, GigaScience (2020).

[8] A. F. Tygel, Semantic tags for open data portals: Metadata enhancements for searchable open data, Federal University of Rio de Janeiro (2016).

[9] M.-H. Hsu, H.-H. Chen, Efficient and effective prediction of social tags to enhance web search, Journal of the American Society for Information Science and Technology (2011).

[10] V. Oliveira, G. Gomes, F. Belém, W. Brandao, J. Almeida, N. Ziviani, M. Gonçalves, Automatic query expansion based on tag recommendation, in: Proceedings of the 21st ACM international conference on Information and knowledge management, 2012.

[11] B. Hong, Y. Kim, S. H. Lee, An efficient tag recommendation method using topic modeling approaches, in: Proceedings of the International Conference on Research in Adaptive and Convergent Systems, 2017.

[12] R. Krestel, P. Fankhauser, W. Nejdl, Latent dirichlet allocation for tag recommendation, in: Proceedings of the third ACM conference on Recommender systems, 2009.

[13] W. Huang, S. Kataria, C. Caragea, P. Mitra, C. L. Giles, L. Rokach, Recommending citations: translating papers into references, in: Proceedings of the 21st ACM international conference on Information and knowledge management, 2012.

[14] B. Sigurbjörnsson, R. Van Zwol, Flickr tag recommendation based on collective knowledge, in: Proceedings of the 17th international conference on World Wide Web, 2008.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013.

[16] K. M. Ting, Precision and Recall, 2010.